Computational Linguistics and Chinese Language Processing
Vol. 18, No. 4, December 2013, pp. 81-96

使用語音評分技術輔助台語語料的驗證

Using Speech Assessment Technique for the Validation of Taiwanese Speech Corpus

李毓哲*、王崇喆*、陳亮宇+、張智星#、呂仁園‡

Yu-Jhe Li, Chung-Che Wang, Liang-Yu Chen,

Jyh-Shing Roger Jang, and Ren-Yuan Lyu

摘要

本論文的主要研究為使用語音辨識及結合語音評分，對未整理的台語語料進行初步的篩選。藉由機器先過濾掉有問題的音檔，如錄音音量過小、太多雜訊、錄音音檔內容有誤等情形，取代傳統人工聽測費時的做法。本論文可分為三個階段，分別是：「基礎聲學模型訓練」、「語音評分與錯誤原因標記」及「效能評估」。

於基礎聲學模型訓練階段，以長庚大學提供的台語語料 ForSD (Formosa Speech Database)為材料，使用隱藏式馬可夫模型（Hidden Markov Model, HMM）進行聲學模型的訓練。聲學模型單位分別為：單音素聲學模型（Monophone acoustic model）、音節內右相關雙連音素聲學模型（Biphone acoustic model）及音節內左右相關三連音素聲學模型（Triphone acoustic model），其針對測試語料進行自由音節解碼辨識網路（Free syllable decoding）的音節辨識率（Syllable accuracy）最佳結果分別為：27.20%、43.28%、45.93%。

*國立清華大學資訊工程學系 Dept. of CS, NTHU, Taiwan
E-mail: {yujhe.li; geniusturtle}@mirlab.org
+國立清華大學資訊與應用研究所 ISA, NTHU, Taiwan
E-mail: davidson.chen@mirlab.org
#國立臺灣大學資訊工程學系 Dept. of CSIE, NTU, Taiwan
E-mail: jang@mirlab.org
‡長庚大學資訊工程學系 Dept. of CSIE, CGU, Taiwan
E-mail: renyuan.lyu@gmail.com

於語音評分與錯誤原因標記階段，將於基礎聲學模型訓練階段已訓練好的左右相關三連音素聲學模型，對待整理的語料進行語音評分，而將其評分結果依照門檻值分為三部分，分別為低分區、中間值區及高分區。且針對低分區部分語料進行人工標記，標記其錯誤原因，再對其擷取特徵，使用支持向量機（Support Vector Machine, SVM）訓練出分類器，最後以該分類器對低分區語料進行二次檢驗，將低分區語料分為可用語料及不良語料。

於效能評估階段，將原先訓練語料分別加入「未整理語料」、「中間值區及高分區語料」、「高分區語料」進行聲學模型的訓練，比較篩選語料前、後效能，其音節辨識率結果分別為：40.22%、41.21%、44.35%。

由結果看來，經過篩選後語料所訓練出的聲學模型與未經篩選語料所產生的聲學模型，其辨識率的差別最高可達 4.13%，證實本論文所提的方法，藉由語音評分確實能有效的自動篩選掉有問題的語句。

關鍵詞：台語語料整理、隱藏式馬可夫模型、語音評分、語音辨識、支持向量機

Abstract

This research focuses on validating a Taiwanese speech corpus by using speech recognition and assessment to automatically find the potentially problematic utterances. There are three main stages in this work: acoustic model training, speech assessment and error labeling, and performance evaluation.

In the acoustic model training stage, we use the ForSD (Formosa Speech Database) ,provided by Chang Gung University (CGU), to train hidden Markov models (HMMs) as the acoustic models. Monophone, biphone (right context dependent), and triphone HMMs are tested. The recognition net is based on free syllable decoding. The best syllable accuracies of these three types of HMMs are 27.20%, 43.28%, and 45.93% respectively.

In the speech assessment and error labeling stage, we use the trained triphone HMMs to assess the unvalidated parts of the dataset. And then we split the dataset as low-scored dataset, mid-scored dataset, and high-score dataset by different thresholds. For the low-scored dataset, we identify and label the possible cause of having such a lower score. We then extract features from these lower-scored utterances and train an SVM classifier to further examine if each of these low-scored utterances is to be removed.

In the performance evaluation stage, we evaluate the effectiveness of finding problematic utterances by using 2 subsets of ForSD, TW01, and TW02 as the

training dataset and one of the following: the entire unprocessed dataset, both mid-scored and high-scored dataset, and high-scored dataset only. We use these three types of joint dataset to train and to evaluate the performance. The syllable accuracies of these three types of HMMs are 40.22%, 41.21%, 44.35% respectively.

From the previous result, the disparity of syllable accuracy between the HMMs trained by unprocessed dataset and processed dataset can be 4.13%. Obviously, it proves that the processed dataset is less problematic than unprocessed dataset. We can use speech assessment automatically to find the potential problematic utterances.

Keywords: Taiwanese Corpus Validation, Hidden Markov Model, Speech Assessment, Support Vector Machine.

1. 緒論

傳統語料的整理往往需要耗費相當的時間以及需要具有專業知識背景的人員進行人工聽測。本論文的研究即是針對台語語料使用語音評分輔助機器篩選掉不良的語料，如空白音檔、文本有誤...等錯誤類型，取代傳統費時的人工聽測方法，藉此減少人工檢查的時間，加快語料庫的建立。

本論文首先會進行基礎聲學模型的訓練，藉由基礎聲學模型對未整理的語料進行語音評分，並依照分數門檻值將未整理的語料劃分為低分區語料、中間值區語料及高分區語料，最後分別將未經篩選語料與經篩選語料加入原始訓練語料重新進行聲學模型的訓練，以測試語料的辨識結果來評估語料整理的程度。

而對於低分區語料，我們特別對它進行人工標記，標記該音檔的不良類型，並依觀測到的不良類型對其擷取特徵，使用支持向量機（Support Vector Machine, SVM）訓練分類器。最後我們則使用該分類器對低分區語料再次檢驗為可用語料或是不良的語料，減少需要人工檢驗的語料數目。底下圖 1 為語料整理系統流程圖：

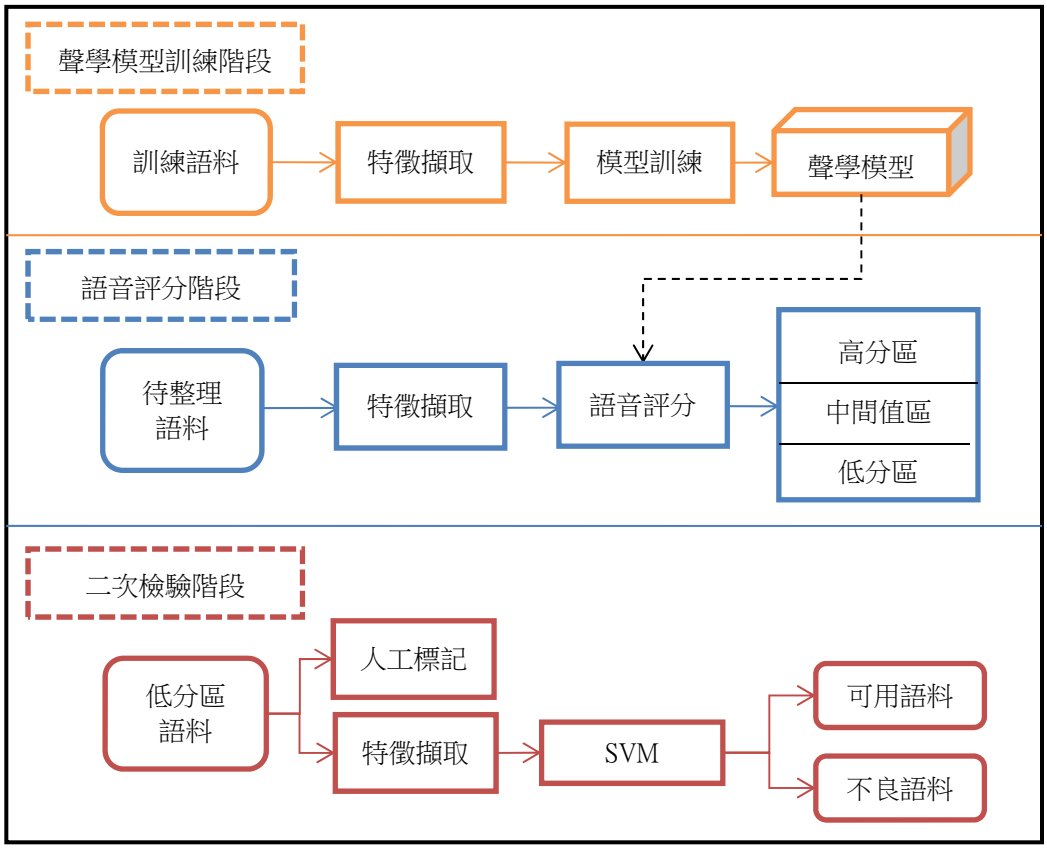

圖 1. 語料整理系統流程圖

因此，本論文的研究方向於如何使用電腦為輔助工具，對未整理語料藉由語音評分進行初步的篩選，並標記出不良類型，且以觀測到的不良類型為特徵產生分類器，對語料進行第二階段的檢驗。以下第二節將介紹論文之相關研究。第三節將詳述論文方法，提出使用語音評分輔助語料的驗證方法。第四節為本論文所提出的方法進行實驗，並對實驗的結果進行探討。第五節為本論文的結論以及未來研究方向。

2. 相關研究

2.1 標音系統與模型訓練

本論文內採用的標音系統為福爾摩沙音標（Formosa Phonetic Alphabet, ForPA）。ForPA 拼音是台灣地區華語、台語、客語三語共用的標音系統，在華語的音素有 37 個，台語有 56 個，兩種語言音素聯集共有 63 個，交集共有 32 個（朱晴蕾等，2010）。進行聲學模型之訓練時，梅爾倒頻譜係數（Mel-frequency Cepstral Coefficients, MFCCs）（Davis & Mermelstein, 1980）和對數能量（Log energy）做為語音特徵；並以隱藏式馬可夫模型（Hidden Markov Model, HMM）來建立不含聲調的聲學模型。

2.2 語音評分

語音評分（Speech assessment）（李俊毅，2002；陳宏瑞，2011）能藉由聲學模型對錄音進行評分，在本論文中以評分後的分數來評量音檔的與文本間的相似程度。但依據前人研究（黃武顯，2007），在某些狀況下語音評分的分數並不合理，因此在本論文中加入了三種分數調整的扣分機制，以降低評分時不合理情形出現的機率。

音節之音框個數差距過大:正常情況下,每個音節所占有的音框個數差距不會太大。當語句中所占的音框個數最長的音節與所占的音框個數最短的音節，音框個數差距達 3 倍以上時，則將分數減少為原先分數的 80%。

音節中連續音素之音框數目過小：在正常情況下，很少出現一個音節中連續數個音素其音框各數皆為 3 個音框，相當於音素的最小音框個數。當一個音節中出現連續兩個音素的音框數皆為 3 個音框，則將分數減少為原先分數的 80%。

音節數目不一：在本論文中我們允許評分時出現漏字的情形，即可跳過某個音節，而不強制對文本內的所有音節都做強迫對位的動作，因此系統評分的切音結果可能會出現缺字的情況。而在（李俊毅，2002; 陳宏瑞，2011）當中所提到的分數計算方法，僅針對切音的結果來進行分數的計算，並不會考慮到輸入語句的所有音節。當經評分後之切音結果的音節數目小於正確文本的音節數目時，則將分數依照切音結果的音節數目與輸入語句之音節數目之比例來調整。例如，原先輸入語句為「如果咱 e 先賢無 e 話 /ru-ger-lan-e-sen-hen-bher-e-ue 」，共 9 個音節，但切音結果後的語句卻為「ger-lan-e-sen-hen-bher-e-ue」,共 8 個音節,因此若調整前為 80 分,則調整後為 80×8/9=71 分。

3. 研究方法

3.1 基礎聲學模型訓練

本論文使用 HTK（Hidden Markov Model Toolkit）（Young, 2009）訓練聲學模型和調整特徵參數。於基礎聲學模型訓練階段，使用 HTK 對訓練語料擷取特徵後，分別訓練出三種不同音素單位的聲學模型，包含了單音素（Monophone）、音節內右相關雙連音素（Biphone）及音節內左右相關三連音素（Triphone）。

3.2 語音辨識結合語音評分

本論文中使用 HMM 聲學模型與語音評分及扣分機制作結合，藉由評分後的分數，當作音檔內容與文本內拚音內容的相似性參考。如果說錄音音檔的品質越好的話，藉由穩定的聲學模型便可以辨識出音檔的語句內容，評分分數也會較高；反之，如果錄音音檔的品質不好，例如有雜訊、錄音音量太小、音檔內容與實際文本不同、音檔片段切音沒切好...等，則會增加辨識的困難度，相對的評分分數會偏低。因此我們將評分分數的高低做為該音檔是否為優良語料的參考依據。

在對待整理語料進行語音評分的方面，首先我們以（李俊毅，2002; 陳宏瑞，2011）

當中的方法，算出該語句的評分分數，再使用二-（二）所提到的扣分機制，產生該語句最終的評分分數。依據不同的分數門檻值，我們將評分後的語料分為低分區、中間值區及高分區。其中低分區內的語料普遍為不良的語料，但仍然可能存在著因為分數誤被低評的可用語料，因此我們藉由人工觀察，更深入地去分析低分區語料的錯誤類型；而依據評分的結果，高分區內的語料具有相當的公信力是屬於較優良的語料，或許當中也有可能出現該錄音音檔的評分分數被高評的的情況，但是因為有扣分機制的加入，該狀況會是極少數。而中間值區的語料屬於比較模糊的區域。

3.3 低分區語料二次檢驗

由於低分區內的語料可能因為語音評分及扣分機制的誤判，造成可用語料的分數被低評。因此本論文提出使用二次檢驗的方法，藉由 SVM 產生的分類器再次檢驗低分區語料，將語料分類為可用的語料和不良的語料。

由於我們事先不知道該語料是否為好的語料但分數被低評或者是不良語料，因此我們需要先對部分的低分區語料進行人工標記，標記其音檔問題，並將已標記的語料分為訓練語料及測試語料。接著再以標記過程中語料常見的錯誤類型當作特徵，藉由參數的調整產生出分類效果不錯的分類器。最後針對低分區內的所有語料使用該分類器進行二次檢驗，並將其被歸類為可用的語料進行人工檢驗。

藉由使用分類器進行二次檢驗的動作可降低誤刪除可用語料的機率，同時不需要針對低分區內的所有語料進行人工檢查，只需對過濾後的可用語料作人工檢驗即可，藉此減少人工檢查時間，加快語料庫的建立。

4. 研究結果與分析

4.1 語料簡介

本論文之語料來源為長庚大學於 2001 至 2003 年間，執行國科會委託計畫「台灣地區多語語音辨認之研究暨多語語音資料庫之建立」所蒐集的台語語料庫 ForSD (Formosa Speech Database)，本論文使用其中的 TW01,TW02,以及 TW03 等三個子集合（Lyu *et al*., 2004）。該語料分為訓練語料與測試語料兩個集合，本論文主要也以語料中分配好的集合做為訓練與測試。表 1 為該語料的相關數據。

表 1. 語料資訊

	訓練語料	測試語料
語料名稱	ForSD-TW01、ForSD-TW02	ForSD-TW01、ForSD-TW02
錄音格式	單聲道，16kHz，16bits	單聲道，16kHz，16bits
錄音者	600 人，男 317 人、女 283 人	26 人，男 13 人、女 13 人
錄音句數	117047 句，男 61908 句、女 55139 句	3072 句，男 1549 句、女 1523 句
錄音時間	32.58 小時	0.98 小時

本論文待整理語料來源為 ForSD 語料中的 TW03 子集合（廖子宇等，2012）。該語料依照錄音者身份可分為老師及學生語料，共由 4 位老師和 671 位學生錄製而成。其中男生錄音人數少於女生錄音人數，比例約為 1:1.5。錄音內容為文章段落內的句子。表 2 為待整理語料的相關數據。

表 2. 待整理語料資訊

	待整理語料
語料名稱	ForSD-TW03
錄音格式	單聲道，16kHz，16bits
錄音者	675 人，男 263 人（老師 2 人、學生 261 人）、女 412 人（老師 2 人、學生 410 人）
錄音句數	205311 句，男 82099 句、女 123212 句
錄音時間	136.14 小時

4.2 辨識網路與效能評估

本論文採用自由音節解碼（Free syllable decoding）做為辨識網路，其中欲辨識音節的個數為 853 個，而音節至下個音節的對數機率值為-20。該辨識網路前後音節設定為 silence，中間為欲辨識的音節，如圖 2 所示。

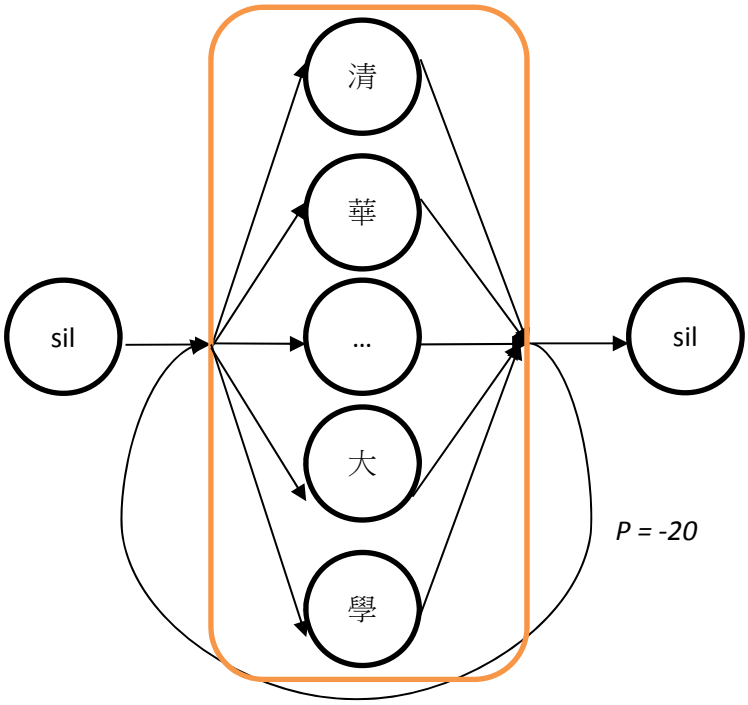

圖2. 辨識網路結構

正常情況下，當訓練時使用的語料量越多，則能訓練出較穩定的聲學模型，辨識效果也會較佳。但如果訓練語料內夾雜著不良的語料，比如有雜音、音檔與文本內容不符合、有咳嗽聲或是有異常停頓等情形出現時，這些不良的語料則會在訓練聲學模型時對其他的語料產生影響，造成訓練出來的聲學模型較不穩定，對辨識結果會產生影響。因此，在這裡我們使用辨識率來做為語料整理乾淨程度的效能評估方法，對篩選後語料與未經篩選語料所產生的聲學模型之辨識率進行比較。

在本論文中，我們採用音節辨識率（Syllable accuracy）做為語料整理乾淨程度的評估方法。底下為其計算公式：

$$Syllable\ accuracy = \frac{N - D - S - I}{N} \times 100\%$$

其中 N 為正確文本內的總音節數目；D 為遺失的音節數目，指的是出現在正確文本內，但在辨識的結果裡卻沒有辨識出來的音節數目；S 為替換的音節數目，指的是將某一個音節辨識成另一個音節的數目；I 為插入的音節數目，指的是辨識結果中除了正確的音節外，多出了不該出現的音節數目。

4.3 基礎聲學模型訓練

本實驗的目的為使用 ForSD-TW01、ForSD-TW02 訓練語料，藉由參數的調整訓練出穩定、辨識率不錯的基礎聲學模型，以供後續語音評分使用。

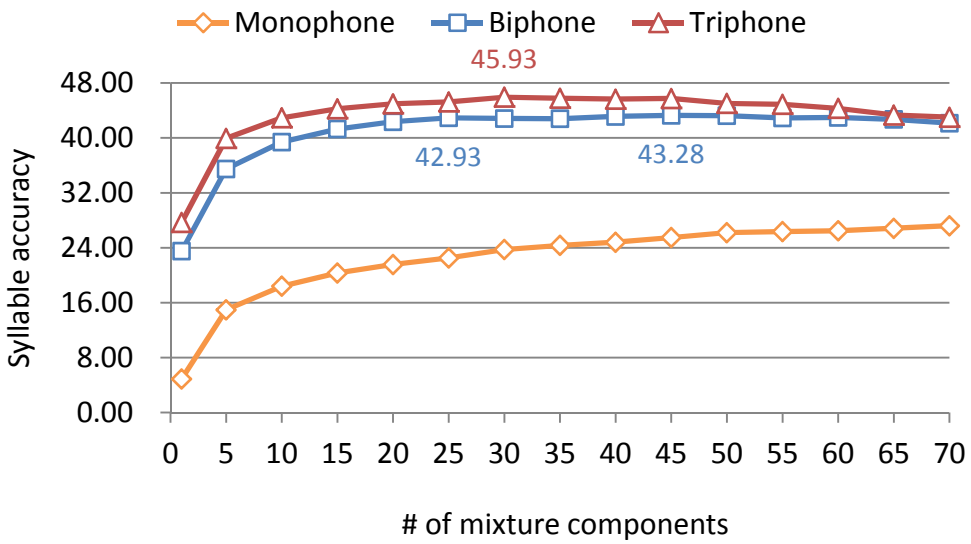

圖 3. Monophone、Biphone、Triphone 聲學模型於不同高斯混和數下之辨識結果

圖 3 列出了 Monophone、Biphone、Triphone 聲學模型於每個狀態下高斯混和數以 5 的倍數，從[1 1 1]至[70 70 70]提升的辨識結果，其中，每個狀態包含三個 Streams，silence 模型的高斯混和數為其他模型的兩倍。

在本實驗中，Triphone 聲學模型於高斯混和數為[30 30 30]時的音節辨識率達到最高，為 45.93%，符合於目前台語語音辨識研究的水平，之後便開始下降；而 Biphone 聲學模型於高斯混和數為[25 25 25]及[45 45 45]時達到高點，之後便開始下降；而 Monophone 聲學模型的音節辨識率則呈現持續上升的趨勢，但明顯與 Biphone、Triphone 聲學模型差距過大。說明了當模型的高斯混和數過高時，則會對該模型過度符合（Over fit），造成辨識率下降。由以上結果，我們選擇高斯混和數為[30 30 30]的 Triphone 聲學模型當作基礎聲學模型，於語音評分階段時使用。

4.4 對待整理語料進行語音評分

本實驗的目的為使用實驗一所產生的基礎聲學模型對待整理語料進行語音評分，並依照分數門檻值將待整理語料分為低分區、中間值區和高分區語料三個區塊。我們首先由前一節所述的基礎聲學模型，對待整裡語料進行強迫對位（Force alignment）及語音評分。在語音評分的過程中，如 2.2 節所述，我們允許忽略文本中欲辨識的音節。得到評分分數後，我們依照分數門檻值將待整理語料分為低分區、中間值區及高分區語料。分數門檻值設定為 60、80 分，將評分分數低於 60 分的語料歸為低分區、評分分數高於 80 分以上的語料歸為高分區、而評分分數為 60 分以上但未達 80 分的語料歸為中間值區。圖 4 為待整理語料進行分區後的句數與時間分布圖，由於一般來說，語者較不會念得特別好或特別壞，所以大部分語料的評分分數落點於 60 至 80 分間。

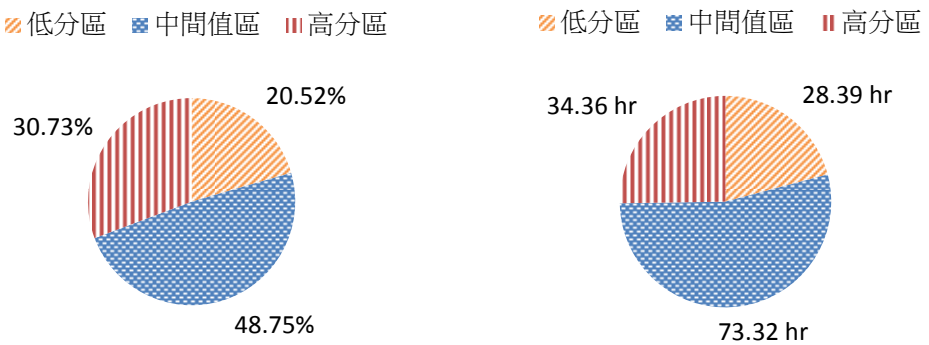

圖 4. 低分區、中間值區、高分區之語料句數（左）與時間（右）分布圖

4.5 未經篩選語料與經篩選語料之聲學模型訓練

本實驗的目的為比較使用未經篩選語料與經語音評分篩選後語料所訓練出的聲學模型，其中使用音節辨識率作為語料整理乾淨程度的評估方法。

本實驗使用上一小節的結果，將原先 ForSD-TW01、ForSD-TW02 訓練語料分別加入未整理語料、中間值區和高分區語料、高分區語料，來進行初始化模型訓練，最終產生三個不同訓練語料組合的聲學模型。於特徵擷取與模型訓練時，使用的基本參數設定皆與基礎聲學模型訓練時相同。

表 3、4、5 分別為加入未經篩選語料、加入中、高分區語料、加入高分區語料所訓練出來的聲學模型其辨識結果。其中，由於模型單位為 Triphone，故語料的多寡對模型數量會稍有影響。從實驗結果可得知，將原始訓練語料加入未整理語料訓練出的聲學模型，其音節辨識率為 40.22%；而加入中間值區及高分區語料訓練出的聲學模型，其音節辨識率為 41.21%；而使用高分區語料訓練出的聲學模型其音節辨識率為 44.35%。此結果比加入待整理語料前的辨識率稍低的原因為，在錄音者及句型等方面，原始訓練語料與測試語料較為接近，但待整理語料與這兩者相差較遠，故將待整理語料加入訓練時，辨識率會稍微降低。

表 3. 未經篩選語料其聲學模型之辨識結果

模型單位	模型數量	總 Mixture 數	Deletion	Substitution	Insertion	Correction	Accuracy
Triphone	1,544	417,150	0.86%	53.04%	5.89%	46.10%	40.22%

表 4. 中間值區及高分區語料其聲學模型之辨識結果

模型單位	模型數量	總 Mixture 數	Deletion	Substitution	Insertion	Correction	Accuracy
Triphone	1,539	415,800	0.82%	51.15%	5.82%	47.03%	41.21%

表 5. 高分區語料其聲學模型之辨識結果

模型單位	模型數量	總 Mixture 數	Deletion	Substitution	Insertion	Correction	Accuracy
Triphone	1,527	412,560	0.95%	49.45%	5.25%	49.60%	44.35%

從實驗結果看來，加入語音評分篩選後的語料所訓練出的聲學模型比加入未經篩選語料所訓練出的聲學模型穩定，其音節辨識率差距可達 4.13%。說明了訓練時使用的語料並非數量越多越好，仍然需要考慮語料的品質。而單純加入高分區語料的辨識效果又比加入中、高分區語料來的好。其原因為中間值區本身屬於一個較模糊的地帶，雖然該區域語料量約為高分區的 1.6 倍，但語料內夾雜不良語料的機率也比高分區高。由此，再次證實了訓練聲學模型時語料品質的重要性。所以經由本實驗，證實了語音評分確實能有效的自動篩選掉有問題的語句。

4.6 低分區語料二次檢驗

由於進行語音評分時，如果使用的聲學模型不夠穩定，則容易產生誤評，造成優良的語料但評分分數卻偏低。本實驗的目的即是使用 SVM 分類器，對低分區的語料進行第二次的檢驗，將低分區的語料分類為可用語料和不良語料。同時藉由前處理人工標記的過程中，歸類出音檔問題的原因。

由於我們事先並不曉得語料為可用語料或是不良語料，因此需要先進行人工標記的工作。標記完成後，再將標記好的檔案切分為訓練檔案和測試檔案，以進行分類器的訓練與效能的評估。

於人工標記階段，我們將低分區內的語料以 10 分為間隔，分別對評分分數為 0 至 10 分、10 至 20 分、...、50 至 60 分的前 250 個音檔進行標記的動作，共標記 1500 個音檔。標記過程中，除了標記為可用語料、不良語料外，同時也標記出歸類為不良語料的原因。標記完成後，以 2 比 1 的比例將已標記好的 1500 個檔案分為訓練檔案和測試檔案。表 6 為訓練檔案、測試檔案的相關數據。其中，標記的 1500 個檔案中有 2 個音檔因為標頭檔損壞，無法使用。

表 6. 低分區二次檢驗分類器之訓練、測試檔案相關數據

	訓練檔案		測試檔案	
標記結果	可用語料	不良語料	可用語料	不良語料
音檔數	246	754	100	398
總音檔數	1,000		498	

於 SVM 分類器訓練階段，我們使用 LIBSVM（Lin, 2013）為工具進行分類器的訓練。於特徵擷取時，我們以人工標記過程中觀察到的不良原因為參考，分別取音檔前、後五個音框的音量（1~10 維）、音檔前、後五個音框音量的一階微分（11~20 維）、語音評分分數（21 維）、依語音評分結果，語句內最長音節和最短音節的比例（22 維）、依語音評分結果，切音結果音節數與文本音節數的比例（23 維）、依語音評分結果，音節的平均音量（24 維）、依語音評分結果，每單位時間內的音節數目（25 維），共 25 維特徵。表 7 為 SVM 分類器的參數設定。

表 7. SVM 分類器參數設定

參數項目	**設定內容**
Cross validation	5-fold
Kernel function	RBF kernel
C	8.0
γ	0.125
Feature dimension	25 dimension

於人工標記的過程中，我們將不良語料其不良的原因歸為 14 種類型，其中一個音檔內可能出現多種錯誤類型。表 8 列出了 1500 個標記檔案的不良原因。由該統計結果看來，標記的音檔內最常見的標記為可用語料、音檔片段切音有誤、音檔音節數目小於文本音節數目、有雜音...等。

表 8. 標記音檔的不良原因及出現次數

不良原因	出現次數(百分比)	不良原因	出現次數(百分比)
可用語料	346 (23.07%)	文本和聽打結果不同	80 (5.33%)
音檔片段切音有誤	333 (22.20%)	音節發音錯誤	62 (4.13%)
音檔音節數小於文本音節數	231 (15.40%)	有爆破音	41 (2.73%)
有雜音	177 (11.80%)	說話速度過快	18 (1.20%)
空白音檔	166 (11.07%)	異常停頓	12 (0.80%)
音檔內容與文本內容不符合	125 (8.33%)	音檔有問題	7 (0.47%)
音量過小	90 (6.00%)	有笑聲	3 (0.20%)

底下列出了上述不良原因的詳細說明：

A. 可用語料：標記為可用語料的語料其音檔內容與文本沒有異狀，為可用的語料。圖 5 為文本內容為「我答應了/ ghua-da-ing-liau」的波形圖，該錄音音檔的內容與文本內容符合，但其評分分數卻低於 60；而造成其低分的原因為音素強制對位時有誤，產生誤評。

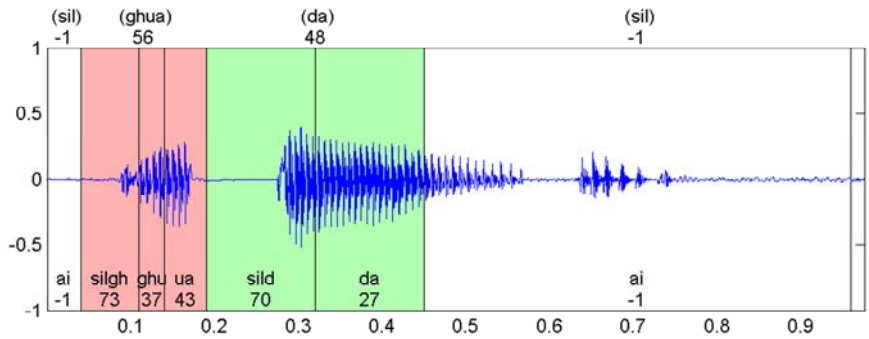

圖 5. 標記為可用語料之波形圖

B. 音檔片段切音有誤：於標記錄音音檔片段的開頭與結尾時，音檔的開頭或結尾恰好切割在音節的中間，造成音檔內前、後音節發音的不完整。以圖 6 為例，該錄音音檔內容為「著會凍真真正正 e 組織/ dier- e- dang- zin- zin- ziann- ziann- e- zo- zit」，但在標記該錄音音檔片段的開頭時，卻切割在音節 dier 中間，造成部分音素的發音遺失。

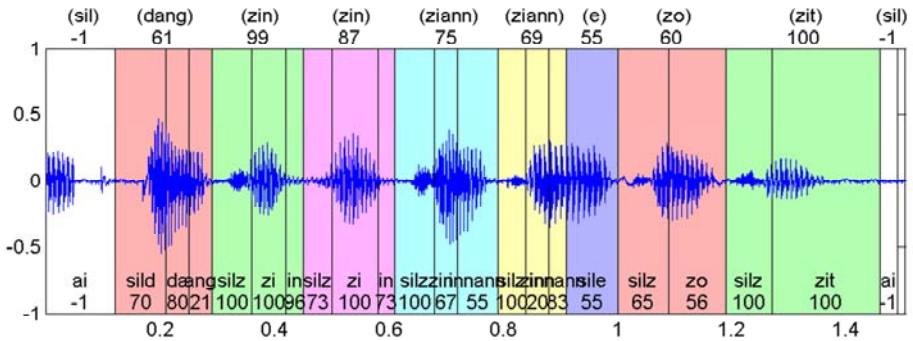

圖 6. 音檔片段切音有誤之波形圖

C. 音檔音節數小於文本音節數：文本內部分音節在錄音音檔內沒有發音，例如文本內容為「有一寡人出門遊覽/u-zit-gua-lang-cut-mng-iu-lam」，但音檔內容卻只發出「寡人出門遊覽/ gua-lang-cut-mng-iu-lam」，少了「有一/u-zit」兩個音節。

D. 有雜音：錄音音檔內，人聲裡頭夾雜著雜音。

E. 空白音檔：錄音音檔內沒有人聲，只有爆破音或純雜音。

F. 音檔內容與文本內容不符合：錄音音檔的內容和真正文本內容不符合，例如文本內容為「東吳大學嘛 m 是東湖大學 dang-gho-dai-hak-ma-m-si-dang-o-dai-hak」，但音檔內容為「東吳大學嘛 m 是嘛 m 是東湖大學/dang-gho-dai-hak-ma-m-si-ma-m-si-dang-o-dai-hak」，多了「嘛 m 是/ma-m-si」三個音節，與文本內容部分不符合。

G. 音量過小：錄音音檔的音量過小。

H. 文本和聽打結果不同：文本的內容和聽打的內容不同，例如文本內容為「殘殺 in-e 族人」，但聽打內容為「zan-sat-zok-rin」，沒拚出「in-e」兩個音節；而錄音音檔又為依據文本內容錄製而成，造成音檔內容和聽打結果對應上有問題。

I. 音節發音錯誤：錄音音檔內部分音節的發音有誤。

J. 有爆破音：錄音過程中，麥克風出現爆音狀況。

K. 說話速度過快：錄音音檔內的說話速度過快，人耳無法辨識所錄製的內容。

L. 異常停頓：錄音音檔內出現異常的停頓，或是某個音節的尾音拖的特別長。

M. 音檔有問題：錄音音檔的標頭檔有問題，或是格式有誤，造成無法讀取音檔。

N. 有笑聲：錄音音檔內夾雜笑聲。

使用由訓練檔案進行 5-fold CV（Cross validation）所得到的最佳參數：C=8.0、γ =0.125，其分類器對測試檔案分類結果的辨識率（Accuracy）為 82.33%。底下為辨識率的計算公式：

$$Accuracy = \frac{\text{分類正確檔案數目}}{\text{測試檔案總數目}} \times 100\%$$

表 9 為本實驗中分類器對測試檔案的混淆矩陣（Confusion matrix）。從表中我們可以求出該分類器針對找出不良語料為目的的錯誤接受率（False Accept Rate, FAR）為 70%；錯誤拒絕率（False Reject Rate, FRR）為 4.52%。以剔除不良語料為目的的狀況下，使用該分類器確實能有效的篩選掉不良語料。

表 9. 混淆矩陣

		預測值	
		可用語料	不良語料
實際值	可用語料	30	70
	不良語料	18	380

藉由 SVM 分類器，我們可以對低分區內所有的語料進行分類，將語料分類為可用語料和不良語料，其分類結果可用語料為 8570 個、不良語料為 33398 個。由實驗的結果，我們只需對被分類為可用語料的音檔進行人工檢驗，從中挑選出實際可用的語料。

同時，我們使用該分類器對高分區語料進行分類，發現被歸類為不良語料的音檔內容，確實也有出現上述人工檢驗觀察到的不良原因。由此結果，說明了我們能藉由該分類器觀察到語音評分無法找出的錯誤。因此，我們可更進一步的對被歸類為不良語料的音檔進行人工檢驗，將有問題的語料剔除。

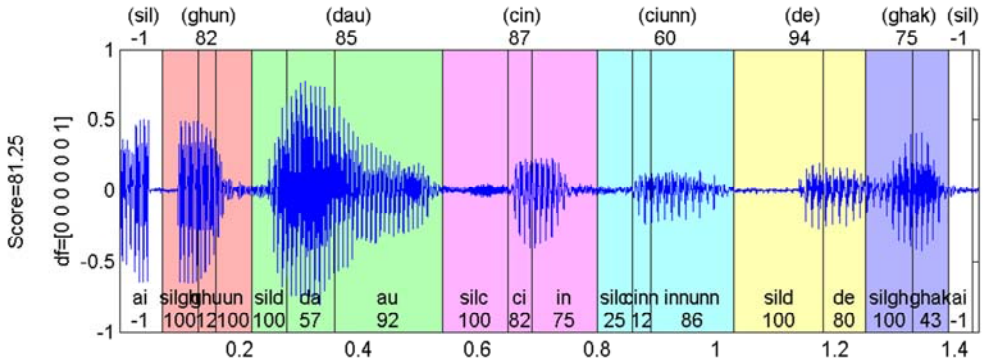

圖 7. 高分區語料音檔片段切音有誤

圖 7 為高分區語料音檔片段切音有誤範例，該錄音音檔的文本內容為「阮兜親像地獄/ghun-dau-cin-ciunn-de-ghak」，語音評分分數為 81.25，但在切割音檔片段時，第一個音節「阮/ghun」的部分音素卻被切掉，造成音節不完整。

5. 結論與未來研究方向

5.1 結論

本論文藉由參數的調整，使用效果最佳的基礎聲學模型對未整理的語料進行語音評分，並依照評分後的分數將語料分為低分區、中間值及高分區，其中我們對低分區的語料進行更深入的觀察。

於未經篩選語料與經篩選語料之聲學模型訓練的實驗結果，加入未篩選語料所產生的聲學模型，其音節辨識率為 40.22%；而加入篩選後語料所產生的聲學模型，其音節辨識率最佳為 44.35%；其辨識率的差別可達 4.13%。證實了藉由語音評分確實能有效的自動過濾掉有問題的語句。

而在觀測低分區語料過程中，我們將語料的不良類型歸類為 14 種，並以語料不良原因為特徵，進行 SVM 分類器的訓練，其辨識效果為 82.33%；我們能利用該分類器對低分區語料再挑選出可用語料，對中、高分區語料再剔除掉不良語料，降低語音評分誤判的機會。同時，由於該分類器僅考慮語料不良的原因，並沒有針對特定語言，因此可適用於各種語言做語料檢測使用。

5.2 未來研究方向

在語音評分部分，聲學模型穩定的程度會影響評分的結果。因此，如何在基礎聲學模型訓練階段，藉由參數的調整產生辨識效果較好的聲學模型會是要點。或許可以試著將待整理語料切成數分，依序對切割的語料進行語音評分、剔除不良語料、加入訓練語料進行聲學模型訓練，並反覆此一過程。

在效能評估部分，目前的測試語料為 ForSD-TW01、ForSD-TW02 測試語料，其中內容多為短詞，而 ForSD-TW03 語料多為長句。可以試著從高分區語料抽取部分語料加入測試語料，增加效能評估的客觀性。

在語料整理部分，雖然 ForSD-TW01 與 ForSD-TW02 是經過人工整理的語料，但也可以利用此研究解果，對語料進行重新分析，以得到更穩定的訓練結果。

致謝

本論文經費來源由國科會計畫 NSC 99-2221-E-007 -049 -MY3，以及 NSC 99-2221-E-182-029-MY3 所提供。

參考文獻

Davis, S. B., & Mermelstein, P. (1980). Comparison of Parametric Representation for Monosyllabic Word Recognition in Continuously Spoken Sentences, *IEEE International Conference on Acoustics*, 1980.

Lin, C.-J. (2013). http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html, 2013

Lyu, R.-y., Liang, M.-s., & Chiang, Y.-c. (2004). Toward Construction A Multilingual Speech Corpus for Taiwanese (Min-nan), Hakka, and Mandarin, *International Journal of Computational Linguistics and Chinese Language Processing*, *9*(2), 1-12.

Young, S. (2009). The HTK Book version 3.4, Microsoft Corporation, 2009.

朱晴蕾、呂道誠、呂仁園(2010)。混合語言之語音的語言辨認。*ISCSLP*。

李俊毅(2002)。*語音評分*。清華大學研究所碩士論文，新竹。

陳宏瑞(2011)。*使用多重聲學模型以改良台語語音評分*。清華大學研究所碩士論文，新竹。

黃武顯(2007)。*基於 32 位元整數運算處理器之華語語音評分的改良與研究*。清華大學研究所碩士論文，新竹。

廖子宇、呂仁園、高明達、江永進、張智星(2012)。台語文字與語音語料庫之建置，*ROCLING 2012*，102-111。